Replication of Flesch et al., 2018
"Comparing continual task learning in minds and machines"
PNAS 115 (44), E10313-E10322.
https://www.pnas.org/content/115/44/E10313

The original paper includes several studies. We randomly chose experiment 1a. In this between-subject experiment, participants perform a "virtual gardening task" where they in 400 training trials and then 200 test trials plant trees in two gardens (north or south). Via trial and error they learn which type of tree grows best in each garden; the trees vary in two dimensions (leafiness and branchiness). Participants are randomized to treatments where they receive training under different conditions. In the "Interleaved group", north and south gardens vary randomly from trial to trial in the training trials; in the "B200 group", gardens are held constant in blocks of 200 trials during the training trials; in the "B20 group", gardens are held constant during blocks of 20 trials during the training trials; and in the "B2 group", gardens are held constant in blocks of 2 trials. We focus on the comparison between the Interleaved group treatment and the B200 group treatment. Participants in the B200 group perform better on the main task than participants in the Interleaved group.

Hypothesis to replicate and bet on: Blocked training (where one of two task rules is held constant for large blocks of trials) leads to higher performance on a subsequent interleaved task (where the two task rules vary randomly from trial to trial) than interleaved training. The authors test the above hypothesis in an independent samples *t*-test ($t(93) = 2.32$, $p = 0.023$); Fig. 2A and p. 10315.

Criteria for replication: The criteria for replication are an effect in the same direction as the original study and a *p*-value < 0.05 in a two-sided independent samples *t*-test.

Power analysis: The original sample had 95 participants in the two treatments. The standardized effect size (Cohen's *d*) was $d = 0.476$. To have 90% power to detect 67% of the original effect size, a sample size of $n = 418$ is required.

Sample: Only participants from the US were allowed to participate in the original study, and only participants who accurately planted >55% of their trees in the training trials were included (where accuracy is measured from whether leafy or branchy trees were planted in north or south). We will use the same inclusion criteria in our replication. We will make sure that participants can only participate once from the same account in this specific study, and we will only recruit participants with a HIT approval rate of 95% or higher. We will also check all IP addresses via https://www.ipqualityscore.com/; and we will remove any participants where one or more of the following is true: fraud score >= 85; TOR = True; VPN = True; Bot = True; abuse velocity = high. The replication sample size is the sample size after any exclusions of participants.

Materials: We will use the same material as the original study, provided in a *Github* repository (https://github.com/summerfieldlab/Flesch_etal_2018) by the original authors. In particular, the experiment will be conducted using a JavaScript platform in forced-fullscreen mode.

Procedure: We will closely follow the procedure of the original study. We will only replicate the Interleaved and B200 treatments and not the other treatments included in the original study. The following summary of the experimental procedure is therefore largely based on the

description of the experiment in the article's Methods section (p. E10321) and the Supplementary Information (pp. 2-3).

Participants will first be shown a Captcha, and will thereafter provide informed consent. After this we will include an attention check that participants will need to pass to continue to the study. This attention check is in addition to any other potential attention check(s) used in the original study. The experiment will be run in forced fullscreen mode. It begins with instructions followed by a training phase (400 trials) and a test phase (200 trials). In both phases, participants will view a tree in one of two contexts (north and south gardens) and decide whether to plant it or not. They will receive feedback (points) according to how well the trees grow. Rejected trees give zero points. During both training and test phases, the tree's leafiness will determine how well it grows in the north garden, while branchiness will determine its growth in the south garden. The number of presentations of each condition (leafy × branchy) is equated in each garden.

Each participant will first face 200 north and 200 south gardens for training. In the B200 group, training gardens remain constant over 200 trials (with the first garden selected at random). In the Interleaved condition, gardens will be randomly selected over trials without replacement. Both conditions will then be evaluated with 100 north and south gardens, respectively, which are interleaved.

Analysis: The analysis will be performed as in the original article. The analysis code was kindly provided by the original authors. In particular, we will perform an independent samples *t*-test comparing the results in the test trials between the B200 group and the Interleaved group.

Subject payments: We are standardizing payments across all replications so that studies have a certain show-up fee depending on the expected length of the study, with an hourly wage from the show-up fee of $8 and a minimum payment of $1 (for studies with incentive payment we use the same incentive payment as in the original study; and this payment is paid in addition to the show-up fee). If we have problems recruiting, we will increase the show-up fee.